第27卷 第3期　工 程 数 学 学 报　Vol. 27 No. 3
2010年06月　CHINESE JOURNAL OF ENGINEERING MATHEMATICS　June 2010

文章编号:1005-3085(2010)03-0389-07

求解高分子结构预测问题的新直接搜索方法*

卢昕玮[1,2]， 张可村[2]

(1- 长安大学经济与管理学院，西安 710064； 2- 西安交通大学理学院，西安 710049)

摘　要: 高分子结构预测问题是目前国际上最热门的研究课题之一，它具有重要的理论意义和实用价值。本文探讨了其中的一类热门模型：势能函数模型。根据目前研究的现状和存在的不足，首次提出了求解该问题的直接搜索算法。首先，本文根据模型的特点改进了筛选子，然后优化了相应的算法。其次，本文进行的数值试验证了该方法的可行性和有效性。与其他方法相比，新算法针对大规模问题的求解具有明显的优势。

关键词: 高分子；势能函数；直接搜索；筛选

分类号: **AMS**(2000) 65K05　　**中图分类号**: O224　　**文献标识码**: A

1 引言

高分子结构预测问题是目前国际上最热门的研究课题之一，它具有重要的理论意义和实用价值，在新物质的发现与研究，新药品的研制和生产，生物信息科学的探索等领域都扮演着十分关键的角色。因此，近几十年来，各国学者在此领域开展了广泛而深入的探索，也取得了令人瞩目的成果。目前，热门的数学模型主要包括下列几个：势能函数[1,2]，距离几何[3]，分子间作用力[4]等。而势能函数模型是其中最受关注的一个。

针对高分子结构预测问题的研究方法主要可归纳为以下三大类：同源建模、折叠识别和从头预测方法。与前两类方法相比，从头预测方法具有相当明显的特点和优势，它不需要序列以外的其他更多信息，仅从一个高分子序列就可以得到相应的空间结构，这是非常理想的理论方法，不仅简洁，还可以发现新的物质结构，这是其他方法所无法匹敌的。

在众多学者的多年努力下，以上三大类方法都取得了不少成果，其中有很多值得一提的好方法，如：Monte Carlo 方法[5]、模拟退火[6]、随机扰动[7]、分支定界[1]及高斯变换[8]等。遗憾的是，现有的求解方法虽然取得了相当的成就但仍然无法满足工程实际中的需求。例如，在求解规模、计算效率、求解精度等方面仍然差强人意。因此，探索高效的新求解方法就变得必要且紧迫。

有鉴于此，本文提出了一个属于从头预测算法类的新直接搜索方法。直接搜索算法[9,10]是一大类不依赖一、二阶导数的方法，最早出现于20世纪50年代，但是它的发展却非常缓慢。随着基于导数算法的兴盛，直接搜索算法被忽视，没有得到足够的重视和重大的发展。然而，随着越来越多的从工业、生物、经济及化学等各个领域中抽象出的大量优化问题都不能用基于导数信息的方法来求解，直接搜索算法以它独特的优势又重新受到了学者的重视，在近年来成为了一个学者研究的热点方向，受到了广泛的关注和重视。直接搜索算法包括的种类很多，主要有模式搜索、线性搜索、共轭方向、二次逼近等。其中模式搜索[11-14]是最热门的一种方法。也是本文探讨和改进的基础。

收稿日期: 2008-03-04.　**作者简介**: 卢昕玮 (1979年12月生)，女，博士，讲师. 研究方向：最优化的理论与应用.
***基金项目**: 国家自然科学基金 (10671057).

全文安排如下：第二部分中首先给出本文要解决的问题的模型。第三部分讨论算法，本文根据模型的特点，首先改进了筛选子，其次给出了改进算法。在第四部分中本文开展了一些数值试验、相应的数据分析和算法比较，得出了新方法更有效等结论。最后一部分是结论和待讨论的问题。

2 势能函数模型

首先，本文给出将要讨论的数学模型，典型的势能函数是根据经典的力学模型结合一定的光谱试验数据等发展出来的。无论是简单或复杂的模型通常包括以下四项[1,15]：键伸缩能、键角变形能、键的转动能(即二面角能)和非键相互作用。用特定的符号转化为数学语言后可以写成

$$\begin{aligned} E &= E_1 + E_2 + E_3 + E_4 \\ &= \sum_{\text{bonds}} \frac{k_i}{2}(l_i - l_{i,0})^2 + \sum_{\text{angles}} (\theta_i - \theta_{i,0})^2 + \sum_{\text{torsions}} c_{ij}^2\big(1 + \cos(3\omega_{ij} - \omega_{ij}^0)\big) + \sum_{i,j} \frac{(-1)^i}{r_{ij}}, \end{aligned}$$

其中 E 代表势能函数，r_{ij} 是不同原子之间的距离。第一项为键伸缩能，即键长 l_i 偏离平衡位置 $l_{i,0}$ 时的能量增量；第二项为键角变形能，即键角 θ_i 偏离平衡位置 $\theta_{i,0}$ 时的能量增量；第三项为二面角项，代表沿着一个给定的键旋转时引起的二面角畸变的能量，它在本质上是周期的，c_{ij} 为一个常数，ω_{ij}^0 是一个参考角；第四项是非键相互作用力，一般都使用势能项。

遗憾的是，尽管上述数学模型已经作了简化和提炼，但是仍然难以求解。因此，在确保原模型仍然能反映实际情况的前提下，有必要进一步优化上述模型。事实上，在以上四个分量中，代表键的转动能 E_3 (即二面角能)和非键相互作用 E_4 在模型的构建中发挥着更重要的作用，即 E_1 和 E_2 项可以忽略。由此可做出假设，$l_i = l_{i,0}$，$\theta_i = \theta_{i,0}$，$i = 1,\cdots,n$，即 E_1 和 E_2 已经处于平衡位置，故 $E_1 = 0$，$E_2 = 0$。这样方程就写成为

$$E = E_3 + E_4 = \sum_{\text{torsions}} c_{ij}^2\big(1 + \cos(3\omega_{ij} - \omega_{ij}^0)\big) + \sum_{i,j} \frac{(-1)^i}{r_{ij}},$$

其中扭转角的余弦为

$$\cos\omega = \frac{\cos\gamma - \cos\alpha\cos\beta}{\sin\alpha\sin\beta}.$$

又

$$\cos\gamma = \frac{r_{ij}^2 + r_{jl}^2 - r_{il}^2}{2r_{ij}r_{jl}},$$

所以有

$$\cos\gamma = \frac{10.60099896 - r_{il}^2}{4.141720682},$$

又由三倍角定理 $\cos 3\gamma = 4\cos\gamma(\cos\gamma^2 - \frac{3}{4})$。可得

$$E_3 = \sum_{i=1}^{m}\sum_{j=1}^{i-1}\Big[1 + 4\Big(\frac{10.60099896 - r_{il}^2}{4.141720682}\Big)^2 - 3\Big(\frac{10.60099896 - r_{il}^2}{4.141720682}\Big)\Big].$$

为方便起见，将上式写成

$$E(x) = \sum_{i=1}^{n}\Big[1 + 4\Big(\frac{10.60099896 - x_i^2}{4.141720682}\Big)^2 - 3\Big(\frac{10.60099896 - x_i^2}{4.141720682}\Big) + \frac{(-1)^i}{x_i}\Big],$$

其中 $E:\mathbf{R}^n \to \mathbf{R}$, $E \in C^2$。同时，上式中的自变量取值范围也不是任意的，因为在工程实际中，通常会应用一些物理手段和测量设备预先测定出分子间距离的大致范围，例如应用核磁共振方式就是目前国际上普遍采用的一种方法。所以，我们有必要加入相应的约束条件，以使该模型更加完备，切合实际要求。由此就形成了如下的约束优化问题

$$\min E(x) = \sum_{i=1}^{n}\Big[1 + 4\Big(\frac{10.60099896 - x_i^2}{4.141720682}\Big)^2 - 3\Big(\frac{10.60099896 - x_i^2}{4.141720682}\Big) + \frac{(-1)^i}{x_i}\Big],$$

$$\text{s.t.}\ \ x \in X = \big\{\, x = (x_i)_{n\times 1} \in \mathbf{R}^n \mid \mathrm{low}_i \le x_i \le \mathrm{up}_i,\ i = 1,\cdots,n \,\big\},$$

其中 low_i, up_i 分别为 x_i 的上、下界，一般取值范围是 $\mathrm{low}_i = -1$, $\mathrm{up}_i = 1$, $i = 1,\cdots,n$。

3 基于筛选技术的模式搜索算法

考虑如下的优化问题

$$\min_{x\in\Omega} f(x),$$

其中 $\Omega \subset \mathbf{R}^n$, $f(x):\Omega \to \mathbf{R}\cup\{+\infty\}$ 为连续可导函数，但是其导数信息不可得或不可靠。

模式搜索算法由 Box[13] 与 Hooke-Jeeves[14] 在 20 世纪 50 年代末最早提出来的，其主要思想是不借助任何导数信息就能产生一个迭代序列 $x^{(k)}$。在每次迭代时，若迭代点能产生更好的最优值则接受，否则就继续寻找。从几何意义上来说，就是寻找具有较小函数值的“山谷”，力图使迭代产生的序列沿“山谷”向最小值点逼近。

首先，引入一些有用的符号：

λ:	网眼	V:	$\mathbf{R}^n$ 上的基	V_+:	V 上产生的有序正基
v_i:	V_+ 中的元素	$\lvert V_+\rvert$:	V_+ 中元素的个数	F:	筛选子
k:	第 k 次迭代				

在近年的研究中，正基被大量的引入到模式搜索算法的研究中[16]。此处，先给出简略的介绍。正基具有以下的两个主要特征。

1) $\mathbf{R}^n$ 中的任意向量都可以表示成 V_+ 中元素的一个非负组合；

2) V_+ 中的任意子集都不是正基。

则正基 V_+ 的定义如下

$$V^{(k)} = \big\{\, v_i^{(k)} \in \mathbf{R}^n : i = 1,2,\cdots,n \,\big\},$$

每个基 $V^{(k)}$ 必须满足以下条件

$$\big|\det\big[v_1^{(k)}, v_2^{(k)},\cdots,v_n^{(k)}\big]\big| > \tau, \qquad \big\|v_i^{(k)}\big\| \le K, \quad \forall\, i \in 1,\cdots,n,$$

其中 τ 与 K 是独立于 k 的正常数。此外，正基的数量也满足以下关系式

$$n + 1 \le |V_+| \le 2n,$$

2 个最常用的正基是

$$V_+^{(k)} = \Big\{\, v_1^{(k)}, v_2^{(k)},\cdots,v_n^{(k)}, -\sum_{n}^{k} v_i^{(k)} \,\Big\},$$

$$V_+^{(k)} = \big\{\, v_1^{(k)}, v_2^{(k)},\cdots,v_n^{(k)}, -v_1^{(k)}, -v_2^{(k)},\cdots,-v_n^{(k)} \,\big\}.$$

在本文中，我们采用正基$e_1, e_2, \cdots, e_n, -e_{n+1}, -e_{n+2}, \cdots, -e_{2n}$，$e_i$是单位基向量。算法在迭代的过程中，需要不断改变网眼的大小和搜索步长进行计算。但是这样以来，计算负担就会大大增加。对于小规模规划问题不大，但是对于中规模和大规模规划问题，这个缺点就变得十分棘手，甚至会导致求解失败。因此，本文考虑引入筛选技术[17,18]来解决这个问题。由于筛选子的选择是筛选技术中最关键，故本文着重于筛选子的进一步优化和改进。

在模式搜索算法中，目标函数至少都是局部最优值，即它比周围的函数值都要小

$$f(x) \le f(x+\lambda v_i), \quad \forall\, v_i \in V_+, \quad i=1,2,\cdots,|V_+|,$$

由此可以定义如下的筛选函数

$$h(x)=\begin{cases} 0, & \text{如果}\ \ w(x)\le 0,\\ w(x), & \text{如果}\ \ w(x)>0, \end{cases}$$

其中

$$w(x)=f(x)-\alpha\min\left\{f(x+\lambda v_i),\, i\in\Lambda\right\}-(1-\alpha)f^*.$$

Λ是正基的一个子集，其中的基与搜索方向的夹角均小于$90°$；f^*是当前迭代中的全局最优值；$\alpha\in(0,1)$是一个接近1的常数。

进一步，定义如下筛选子：若对于任意的$x_j\in F^{(k)}$，点均满足

$$h(x)<(1-\zeta)h(x_j) \quad \text{或} \quad f(x)<f(x_j)-\delta h(x), \tag{1}$$

则称该点能被筛选子接受，当有新的点被筛选子接受时，要删除它所控制的所有点[19]。其中$\zeta,\delta\in(0,1)$是接近0的参数。

本文改进的筛选子具有如下的优势：

1) 既提高了搜索能力又确保了较快的搜索速度。一方面，搜索的范围有效扩大了，只要能使$f(x)$或者$h(x)$下降的点都可以作为新的迭代点；另一方面，本文也舍弃了那些不太可能出现最小点的搜索方向。

2) 提高了方法求解大规模问题的能力。

下面先引入一个更新准则，若一个点能被筛选子接受，则将其加入到筛选子内，更新筛选子，并删除被该点控制的所有点。算法步骤：

步骤1 初始化$k=1$，α,ζ,δ。令$x^{(1)}\in\Omega$为初始点。$F^{(1)}$为初始筛选子且满足$x^{(1)}\in F^{(1)}$，选择正基$V_+=\{v_1,v_2,\cdots,v_{2n}\}$和初始步长$\lambda^{(1)}$；

步骤2 选择步长$\lambda^{(k)}$，置$i=1$，$p=0$，令$t^{(k)}=|V_+|$；

步骤3 计算点$x^{(k)}+\lambda^{(k)}v_i\in\Omega$，$i\in\Lambda$，上的$f(x)$和$h(x)$函数值；

步骤4 若点能被筛选子$F^{(k)}$接受，则令其为$x^{(k+1)}$，同时根据更新准则更新筛选子$F^{(k+1)}$，扩大步长$\lambda^{(k)}$，$k=k+1$，$p=0$，转步骤6，否则，转步骤5；

步骤5 执行有限步搜索过程，若Ω内存在可被筛选子接受的点，则令其为$x^{(k+1)}$，$p=0$，同时根据更新准则更新筛选子，否则，$p=p+1$；

步骤6 令$i=i+1$，若$i>t^{(k)}$，则$i=1$，若$p<t^{(k)}$，则转步骤2，否则转步骤7；

步骤7 执行有限步搜索过程，若在Ω内存在点可被筛选子接受，则令其为$x^{(k+1)}$，$k=k+1$同时根据更新准则更新筛选子并扩大步长$\lambda^{(k)}$，否则缩小$\lambda^{(k)}$；

步骤8 若停机条件不满足，则转步骤2，否则输出最优值，停机。

说明：

1) 步长 λ 的选择是任意的。但需要注意的是，如果 λ 太小，则很难取得明显的下降，如果 λ 太大，则会降低算法的效率。最通常的选择是搜索成功时 λ 扩大为原来的2倍；失败时缩小为原来的 $\frac{1}{2}$。但是针对不同规模的问题 λ 的选择不尽相同。

2) 步骤5和步骤7中的有限步搜索过程也是任意的。最简单的一种方式就是在可行域内选择随机点。当然，其它更合理的方法也可以使用。

3) 算法的停机标准也是较灵活的。通常的停机标准是 $\lambda^{(k)} < \varepsilon$，即当步长小于某一预定值时，算法停机。

收敛性证明类似于文献[17]，此处略去。

4 数值实验

在本节中，我们通过数值实验来测试新方法的可行性与有效性。设 $\alpha = 0.9$, $\zeta = 0.1$, $\delta = 0.1$, $\varepsilon = 1.00e-05$。作者使用VC++6.0编程，电脑配置为Intel(R) Pentium(R) M1.73G，(768M内存)。主要的结果在表1中列出，并同时列出传统的分支定界方法[1]的计算结果，该方法是求解此模型的一类热门算法。表中PF optimum和PF CPU time (s)是新方法求得的最优值和CPU时间。而BB optimum和BB CPU time (s)则是分支定界方法的最优值和CPU时间。

通过对表1的分析，不难看出新方法可以顺利的求解出所有问题的最优解，并且具有明显的优势，具体表现在以下3个方面：

1) 针对大规模问题的求解具有不可比拟的优势

新方法的最大求解能力达到了2000维，而分支定界方法仅为28维。如果选择适当的参数值 λ，算法的求解能力能够更高。

2) 更高的求解效率

当 $n \leq 50$ 时，新方法所有的CPU计算时间均小于1.0秒。当 $n = 2000$，其CPU计算时间也仅有5008.96秒，约1.39小时。而分支定界的计算时间则大大长于新算法。计算仅28维的问题就花了约82.528小时。此对比充分说明了新法具有更高的求解效率。

3) 更高的计算精度

新方法的计算精度均大于 10^{-1}，且最高可以达到 10^{-4}。但原有的分支定界方法的最高精度仅为 10^{-1}。

综上所述，新方法不仅是可行的，而且是高效的。新方法具有较高的求解精度和效率，同时在求解大规模问题时具有不可比拟的优势。

5 结论与展望

本文探讨了一种求解高分子结构预测问题的新方法。由于现有的很多方法在求解效率、求解规模以及求解精度等方面上不尽如人意，所以亟待探索新思路和方法。本文根据以上存在的问题首次提出了求解该问题的基于筛选技术模式搜索方法。首先本文根据模型的特点改进了筛选子，其次改进了相应的算法。随后进行的数值试验证明了本方法的可行性和有效性。与其他方法相比，新算法在各个方面都有优势，特别是针对大规模问题的求解具有不可比拟的优势。在以后的研究中作者将进一步研究该方法在其他更复杂的模型中的应用。

表 1: 数值实验结果及比较

n	BB optimum	BB CPU time (s)	PF optimum	PF CPU time (s)
10	-5.894e-001	2.47	-2.87862e-003	0.01
11	-3.289e-001	4.32	3.39224e-001	0.02
12	-6.716e-001	6.70	-3.45459e-003	0.02
13	-4.112e-001	12.24	3.79987e-001	0.04
14	-7.539e-001	22.47	-4.03022e-003	0.03
15	-4.934e-001	34.56	3.38073e-001	0.04
16	-8.361e-001	20.53	-4.6059e-003	0.05
17	-5.757e-001	97.20	3.37497e-001	0.05
18	-9.183e-001	218.4	-5.18131e-001	0.06
19	-6.579e-001	448.2	2.71163e-001	0.111
20	-1.0006	1167	9.15127e-001	0.111
21	-7.401e-001	3075	3.8514e-001	0.12
22	-1.0828	6365	-6.333e-003	0.09
23	-8.224e-001	10126	2.61704e-001	0.10
24	-1.1650	19480	-6.90836e-003	0.10
25	-9.046e-001	34657	3.35194e-001	0.12
26	-1.2473	62730	-7.48477e-003	0.13
27	-9.868e-001	123232	2.7311e-001	0.27
28	-1.3295	297100	-8.90219e-002	0.211
30	/	/	-5.41749e-004	0.23
50	/	/	-5.18526e-002	0.631
100	/	/	1.66808e-001	2.544
200	/	/	2.28535e-001	8.182
500	/	/	2.70119e-002	75.178
1000	/	/	8.8819e-001	206.467
1500	/	/	-4.3154e-001	1119.71
2000	/	/	-3.51427e-001	5008.96

参考文献:

[1] Charlile Lavor, Nelson Maculan. A function to test methods applied to global minimization of potential energy of molecules[J]. Numerical Algorithms, 2004, 35: 287-300

[2] Marco Locatelli, Fabio Schoen. Fast global optimization of difficult Lennard-Jones clusters[J]. Computational Optimization and Application, 2002, 21: 55-70

[3] More J J, Wu Z J. Distance geometry optimization for protein structures[J]. J Global Optim, 1999, 15: 219-234

[4] Andrew R Leach. Molecular Modelling Principles and Applications[M]. London: Addison Wesley Longman, 1996

[5] Walse D J, Doye J P K. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms[J]. J Phys Chem, 1997, 101: 5111-5116
[6] Moret M A, Pascutti P G, Bisch P M, *et al*. Stochastic molecular optimization using generalized simulated annealing[J]. J Comut Chem, 1998, 19: 647-657
[7] Zou Z, Richard H Bird, Robert B Schnabel. A stochastic/pertubation global optimization algorithm for distance geometry problems[J]. J Global Optim, 1997, 11: 91-105
[8] Le Thi Hoai An. Solving large scale molecular distance geometry problems by a smoothing technique via the Guassian transform and D C programming[J]. J Global Optim, 2003, 27: 375-397
[9] Kolda T G, Lewis R M, Torczon V. Optimization by direct search: new perspectives on some classical and modern methods[J]. SIAM Review, 2003, 45: 385-482
[10] Lewis R M, Torczon V, Trosset M W. Direct search methods: then and now[J]. J Comput Appl Math, 2000, 124: 191-207
[11] Audet Charles, Dennis Jr J E. A pattern search filter method for nonlinear programming without derivatives[J]. SIAM J Optim, 2004, 14: 980-1010
[12] Coope I D, Price C J. On the convergence of grid-based methods for unconstrained minimization[J]. SIAM J Optim, 2001, 11: 859-869
[13] Box G E P. Evolationary operation: a method for increasing industrial productivity[J]. Appl Stat, 1957, 6: 81-101
[14] Hooke R, Jeeves T A. Direct search solution of numerical and statistical problems[J]. J ACM, 1961, 8: 212-229
[15] 唐焕文，靳利霞，计明军. 蛋白质结构预测的优化模型与方法[J]. 工程数学学报，2002, 19(2): 13-22
Tang H W, Jin L X, Ji M J. Optimization models and algorithms for protein structure prediction[J]. Chinese Journal of Engineering Mathematics, 2002, 19(2): 13-22
[16] Coope I D, Price C J. Frame based methods for unconstrained optimization[J]. J Optim Theory Appl, 2000, 107: 261-274
[17] Wu T, Sun L P. A filter-based pattern search method for unconstrained optimization[J]. Numerical Mathematics, 2006, 15: 209-216
[18] Price C J, Coope I D. Frames and grids in unconstrained and linearly constrained optimization, a nonsmooth approach[J]. SIAM J Optim, 2003, 14: 415-438
[19] Fletcher R, Leyffer S. Nonlinear programming without a penalty function[J]. Math Prog, 2002, 91: 239-269

A New Direct Search Method for Solving Macromolecular Structure Prediction Optimization Problems

LU Xin-wei[1,2], ZHANG Ke-cun[2]

(1- School of Economy and Management, Chang'an University, Xi'an 710064;
2- College of Science, Xi'an Jiaotong University, Xi'an 710049)

Abstract: The macromolecular structure prediction problem is one of worldwide popular research topics with significant theoretical and industrial importance in recent years. In this paper, a modified direct search method based on the filter technique for solving macromolecular structure prediction problems is proposed. First, we modify the filter according to the characteristic of the macromolecular structure model and improve the corresponding approach. Then numerical experiments show the effectiveness of the proposed method. Especially, the proposed approach on solving higher dimensional problems is exceptional.

Keywords: macromolecular; potential energy function; direct search; filter

Received: 04 Mar 2008. **Accepted:** 20 June 2009.
Foundation item: The National Natural Science Foundation of China (10671057).